

1309.43490X00

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicants: K. TAMURA, et al

Serial No.: 10/769,805

Filing Date: February 3, 2004

For: STORAGE SYSTEM AND STORAGE CONTROLLER

LETTER CLAIMING RIGHT OF PRIORITY

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

May 5, 2004

Sir:


Under the provisions of 35 USC 119 and 37 CFR 1.55, applicants hereby claim
the right of priority based on:

Japanese Application No. 2003-393647
Filed: November 25, 2003

A Certified copy of said application document is attached hereto.

Acknowledgement thereof is respectfully requested.

Respectfully submitted,



Carl I. Brundidge
Registration No. 29,621
ANTONELLI, TERRY, STOUT & KRAUS, LLP

CIB/jdc
Enclosures
703/312-6600

日 本 国 特 許 庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日 2 0 0 3 年 1 1 月 2 5 日
Date of Application:

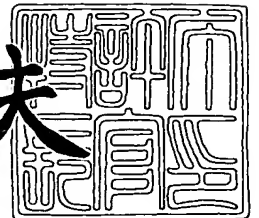
出 願 番 号 特 願 2 0 0 3 - 3 9 3 6 4 7
Application Number:
[ST. 10/C] : [J P 2 0 0 3 - 3 9 3 6 4 7]

出 願 人 株式会社日立製作所
Applicant(s):

2 0 0 4 年 1 月 2 7 日

特許庁長官
Commissioner,
Japan Patent Office

今 井 康 夫



出証番号 出証特 2 0 0 4 - 3 0 0 2 9 5 7

【書類名】 特許願
【整理番号】 340301150
【あて先】 特許庁長官殿
【国際特許分類】 G06F 03/06
【発明者】
 【住所又は居所】 神奈川県小田原市中里 3 2 2 番 2 号 株式会社日立製作所 R A I
 D システム事業部内
 【氏名】 岡本 誉
【発明者】
 【住所又は居所】 神奈川県小田原市中里 3 2 2 番 2 号 株式会社日立製作所 R A I
 D システム事業部内
 【氏名】 長屋 英弘
【特許出願人】
 【識別番号】 000005108
 【氏名又は名称】 株式会社日立製作所
【代理人】
 【識別番号】 100095371
 【弁理士】
 【氏名又は名称】 上村 輝之
【選任した代理人】
 【識別番号】 100089277
 【弁理士】
 【氏名又は名称】 宮川 長夫
【選任した代理人】
 【識別番号】 100104891
 【弁理士】
 【氏名又は名称】 中村 猛
【手数料の表示】
 【予納台帳番号】 043557
 【納付金額】 21,000円
【提出物件の目録】
 【物件名】 特許請求の範囲 1
 【物件名】 明細書 1
 【物件名】 図面 1
 【物件名】 要約書 1
 【包括委任状番号】 0110323

【書類名】 特許請求の範囲**【請求項 1】**

ディスクアレイ装置に用いられ、上位装置との間のデータ授受を制御するチャンネルアダプタであって、

外部メモリからのデータを記憶する内部メモリと、

前記内部メモリに入力される前記データについて入力保証コードを算出し、前記内部メモリから読み出される前記データについて出力保証コードを算出してそれぞれ保持可能な保証コード演算部と、

前記内部メモリから読み出された前記データを前記上位装置に送信する通信部と、

前記データの部分データを前記上位装置に再送信する場合は、前記通信部によって前記部分データを前記上位装置に送信させ、前記出力保証コード演算部により前記データについて再計算された前記出力保証コードと前記入力保証コードとを比較し、これら両保証コードが適合する場合に前記部分データの送信が正常に行われたと判定する制御部と、を備えたチャンネルアダプタ。

【請求項 2】

前記保証コード演算部は、前記内部メモリに入出力されるデータの全体について、前記入力側保証コードまたは前記出力側保証コードを算出するものである請求項 1 に記載のチャンネルアダプタ。

【請求項 3】

前記制御部は、前記部分データを前記上位装置に再送信する場合に、

- (1) 前記通信部により前記部分データを前記上位装置に送信させ、
 - (2) 前記データを前記内部メモリから読み出して前記通信部にダミー転送し、
 - (3) 前記ダミー転送される前記データについての前記出力保証コードを前記保証コード演算部により再計算させ、
 - (4) 前記再計算された出力保証コードと前記データを前記内部メモリに記憶させる際に算出された前記入力保証コードとを比較し、
 - (5) 前記再計算された出力保証コードと前記入力保証コードとが一致する場合は、前記部分データの送信が正常に行われた旨を前記上位装置に通知し、前記再計算された出力保証コードと前記入力保証コードとが不一致の場合は、前記部分データの送信が正常に行われなかった旨を前記上位装置に通知する、
- ものである請求項 1 に記載のチャンネルアダプタ。

【請求項 4】

前記制御部は、前記部分データを前記上位装置に再送信する場合に、

- (1) 部分データを前記内部メモリから読み出して前記通信部に転送することにより、前記部分データを前記上位装置に送信させ、
 - (2) 前記保証コード演算部により、前記内部メモリから読み出された前記部分データについて部分出力保証コードを算出させ、
 - (3) 前記部分出力保証コードと前記データを前記内部メモリに記憶させる際に算出された前記入力保証コードとが不一致であることを確認し、
 - (4) 前記データを前記内部メモリから読み出して前記通信部にダミー転送し、
 - (5) 前記ダミー転送される前記データについての前記出力保証コードを前記保証コード演算部により再計算させ、
 - (6) 前記再計算された出力保証コードと前記データを前記内部メモリに記憶させる際に算出された前記入力保証コードとを比較し、
 - (7) 前記再計算された出力保証コードと前記入力保証コードとが一致する場合は、前記部分データの送信が正常に行われた旨を前記上位装置に通知し、前記再計算された出力保証コードと前記入力保証コードとが不一致の場合は、前記部分データの送信が正常に行われなかった旨を前記上位装置に通知する、
- ものである請求項 1 に記載のチャンネルアダプタ。

【請求項 5】

前記制御部により前記通信部にダミー転送されたデータは、前記上位装置に送信されることなく破棄される請求項 2 または請求項 3 のいずれかに記載のチャンネルアダプタ。

【請求項 6】

前記チャンネルアダプタと前記上位装置との間のデータ授受は TCP/IP プロトコルに従って行われ、前記チャンネルアダプタと前記外部メモリとの間のデータ授受はファイバチャンネルプロトコルに従って行われる、請求項 1 に記載のチャンネルアダプタ。

【請求項 7】

ディスクアレイ装置に用いられ、上位装置との間のデータ授受を制御するチャンネルアダプタの制御方法であって、

前記上位装置への送信済データの部分データを前記上位装置に再送信する場合は、

内部メモリから前記部分データを読み出して前記上位装置に送信させ、

前記データの全体について算出された出力保証コードと前記内部メモリに前記データを入力する際に算出された入力保証コードとを比較し、

これら両保証コードが適合する場合に前記部分データの送信が正常に行われたと判定させるチャンネルアダプタの制御方法。

【請求項 8】

ディスクアレイ装置に用いられ、上位装置との間のデータ授受を制御するチャンネルアダプタの制御方法であって、

外部メモリからのデータを内部メモリに記憶させる第 1 ステップと、

前記内部メモリに前記データを入力する際に、前記データの全体について入力保証コードを算出し保持する第 2 ステップと、

前記内部メモリから前記データを読み出す第 3 ステップと、

前記読み出されたデータの全体について出力保証コードを算出し保持する第 4 ステップと、

前記読み出されたデータを前記上位装置に送信させる第 5 ステップと、

前記入力保証コードと前記出力保証コードとを比較することにより、正常に送信が行われたか否かを判定し、前記上位装置に通知する第 6 ステップと、

前記上位装置に送信された前記データの部分データについて、前記上位装置から再送信要求があったか否かを判定する第 7 ステップと、

前記部分データの再送信要求があった場合は、前記内部メモリから前記部分データを読み出して、前記上位装置に送信させる第 8 ステップと、

前記内部メモリに記憶されたデータを再度読み出す第 9 ステップと、

前記再度読み出されたデータの全体について改めて出力保証コードを算出する第 10 ステップと、

前記改めて算出された出力保証コードと前記第 2 ステップで算出された入力保証コードとを比較することにより、正常に送信が行われたか否かを判定し、前記上位装置に通知する第 11 ステップと、

を含んだチャンネルアダプタの制御方法。

【請求項 9】

上位装置に通信ネットワークを介して接続され、前記上位装置との間のデータ授受を制御するチャンネルアダプタと、

データを記憶する記憶デバイスと、

前記記憶デバイスとの間のデータ授受を制御するディスクアダプタと、

前記記憶デバイスから読出したデータまたは前記上位装置から受信したデータを記憶するキャッシュメモリとを備え、

前記チャンネルアダプタは、

前記キャッシュメモリからのデータを記憶する内部メモリと、

前記内部メモリに前記データを記憶させる場合に、前記データについて入力保証コードを算出して保持する入力保証コード演算部と、

前記内部メモリから前記データを読み出す場合に、前記データについて出力保証コード

を算出して保持する出力保証コード演算部と、

前記内部メモリから読み出された前記データを前記上位装置に送信する通信部と、
制御部とを備え、

前記制御部は、前記データの部分データを前記上位装置に再送信する場合に、

(1) 前記通信部によって前記部分データを前記上位装置に送信させ、

(2) 前記出力保証コード演算部により前記データについて前記出力保証コードを再計算させ、

(3) 前記再計算された出力保証コードと前記入力保証コードとを比較し、これら両保証コードが適合する場合に前記部分データの送信が正常に行われたと判定する、ものであるディスクアレイ装置。

【請求項 1 0】

データ転送制御部を介して内部メモリに接続され、前記内部メモリから読み出したデータを上位装置に送信させる通信部であって、

前記内部メモリに記憶されているデータの部分データを読み出して前記上位装置に送信した後、前記内部メモリから前記データの全体を読み出して、前記上位装置に送信することなく、読み出した前記データを破棄する通信部。

【書類名】 明細書

【発明の名称】 チャネルアダプタ及びディスクアレイ装置

【技術分野】

【0001】

本発明は、ディスクアレイ装置及びチャネルアダプタに関する。

【背景技術】

【0002】

ディスクアレイ装置は、多数の記憶デバイスをアレイ状に配設し、例えば、RAID (Redundant Array of Independent Inexpensive Disks) に基づいて構築された大容量の記憶制御装置である。記憶デバイスとしては、例えば、ハードディスク装置や半導体メモリ装置等が用いられる。各記憶デバイスが提供する物理的な記憶領域上には、論理ボリュームが形成される。業務用サーバ等のホストコンピュータは、所定のコマンドを発行することにより、論理ボリュームからデータを読み出したり、論理ボリュームにデータを書き込むことができる。

【0003】

SAN (Storage Area Network) の一部を構成するディスクアレイ装置は、ファイバチャネルプロトコルに従って、ブロック単位でのデータ転送を行う。ファイバチャネルに従うSANは、FC-SAN等と呼ばれる。FC-SANでは、光ファイバケーブルまたはメタルケーブルを用いて、高速なデータ転送を行うことができる。また、ファイバチャネルスイッチ等を用いることにより、ファブリック構造を得ることができ、多数のディスクアレイ装置と多数のホストコンピュータとを相互に接続することもできる。

【0004】

ところで、ディスクアレイ装置の信頼性を図る1つの観点は、正確にデータの送受信が行えたか否かである。このため、ディスクアレイ装置では、例えば、LRC (Longitudinal Redundancy Check: 水平冗長検査) やCRC (Cyclic Redundancy Check: 巡回冗長検査)、ECC (Error-Correcting Code) 等の技術を採用する。

【0005】

ホストコンピュータとの間のデータ通信を行うチャネルユニットと、磁気ディスクとの間のデータ通信を行うディスクインターフェース制御基板等とのデータ授受が正確に行われたかを検査する技術も知られている (特許文献1)。この技術では、ディスクインターフェース制御基板と、チャネルインターフェース制御基板と、チャネルユニットとの3つの制御基板に、それぞれデータバッファ及びパリティチェック回路を設ける。そして、制御基板間でデータを転送する際には、送り手と受け手の各制御基板でそれぞれパリティチェックを行うと共に、送り手側のデータバッファの記憶内容と受け手側のデータバッファの記憶内容とを比較し、データが正しく転送されたか否かを判定している。

【特許文献1】 米国特許第5561672号明細書

【発明の開示】

【発明が解決しようとする課題】

【0006】

FC (Fibre Channel) - SANにおいて、ディスクアレイ装置からホストコンピュータにデータを送信する場合、例えば2KB程度のブロックデータを含むフレームとして、ホストコンピュータに送信される。ファイバチャネルのフレームは、フレームのスタートを示す情報 (SOF) と、フレームヘッダ情報と、データフィールドと、CRCと、フレームの終了を示す情報 (EOF) とを備えている。データフィールドには、0~2112バイトのデータを格納することができる。ホストコンピュータからの要求に応じて1つまたは複数のフレームが、ディスクアレイ装置からホストコンピュータに送られる。

【0007】

FC-SANは、高速なデータ転送が可能であり、IPネットワークに比べて安定したネットワークである。従って、送信データが途中で消失したり、ホストコンピュータに受信されたデータに回復不能なエラー等が生じる可能性は少ない。これらの事情等から、F

C-SANでは、例えば、送信データの一部に障害等が発生した場合は、送信データの全体を送り直すようになっている。

【0008】

ところで、近年では、インターネットやイントラネット等のようなTCP/IP (Transmission Control Protocol/Internet Protocol) 等をベースにした広域ネットワークシステムが広く普及しており、IP (Internet Protocol) ネットワークとストレージシステム (ディスクアレイ装置) との融合を図る技術が提案されている。IPベースのSANは、IP-SAN等と呼ばれる。IP-SANの一つの形態としては、例えば、SCSI (Small Computer System Interface) のコマンドセットをTCP/IPのパケットとしてシリアル化し、ディスクアレイ装置をIPネットワークに直接接続させる技術が知られている。この技術は、iSCSI (Internet SCSI) あるいはSCSI over IPとして知られている。

【0009】

IPネットワークは、多数のサーバ、ルータ、スイッチ等から構成された複雑なネットワークであり、帯域幅や混雑度も時々刻々と変化する。送り手から相手先に送信されたパケットは、送信順通りに到着する保証はなく、パケットの全部または一部が途中で消失する場合もある。そこで、TCP/IPネットワークでは、一連のデータの一部が消失したり、あるいはエラーが発生しているような場合に、この消失等したデータのための部分的な再送信を要求できるようになっている。

【0010】

IPネットワークにディスクアレイ装置を直結させる場合、ディスクアレイ装置の外部で使用されるプロトコル (TCP/IP) と、ディスクアレイ装置の内部で使用されるプロトコル (FC) との技術的性質の相違が問題となる。即ち、フレーム内のデータの一部についてホストコンピュータから再送信が要求された場合に、要求されたデータそのものは送信することができるが、その部分的なデータが正しいデータであることを保証することができない。ファイバチャネルでは、フレーム全体について1つのCRCを設けるため、フレーム内の一部のデータを再送信する場合、このデータについてデータ保証を行うことができない。

【0011】

本発明は、上記の問題点に鑑みてなされたもので、目的の1つは、上位装置にデータの一部を送信した場合に、この部分的なデータの保証を行うことができるディスクアレイ装置及びチャネルアダプタを提供することにある。本発明の目的の1つは、保証コードが関連付けられる単位よりも小さい量のデータを送信した場合でも、データ保証を行うことができるディスクアレイ装置及びチャネルアダプタを提供することにある。本発明の更なる目的は、後述する実施の形態の記載から明らかになるであろう。

【課題を解決するための手段】

【0012】

上記課題を解決すべく、本発明に従うチャネルアダプタは、ディスクアレイ装置に用いられ、上位装置との間のデータ授受を制御するものであって、内部メモリと、保証コード演算部と、通信部と、制御部とを備えている。内部メモリは、外部メモリからのデータを記憶する。保証コード演算部は、内部メモリに入力されるデータについて入力保証コードを算出し、内部メモリから読み出されるデータについて出力保証コードを算出してそれぞれ保持可能なものである。通信部は、内部メモリから読み出されたデータを上位装置に送信する。制御部は、データの部分データを上位装置に再送信する場合は、通信部によって部分データを上位装置に送信させ、出力保証コード演算部によりデータについて再計算された出力保証コードと入力保証コードとを比較し、これら両保証コードが適合する場合に部分データの送信が正常に行われたと判定する。

【0013】

上位装置は、例えば、ディスクアレイ装置を通信ネットワークを介して利用するコンピュータ装置であり、パーソナルコンピュータ、ワークステーション、サーバマシン、メインフレーム、携帯情報端末等として構成される。内部メモリは、チャネルアダプタの内部

に設けられるもので、例えば、読み書き可能な半導体メモリ（RAM: Random Access Memory）である。内部メモリは、揮発性メモリでも不揮発性メモリでもよい。外部メモリは、チャンネルアダプタの外部に設けられたメモリであり、例えば、キャッシュメモリ等が該当する。上位装置からの要求に応じて、外部メモリから読み出されたデータは、内部メモリに格納され、内部メモリから通信部を介して上位装置に送信される。保証コード演算部は、内部メモリにデータが入力される場合（データ書込み）と内部メモリからデータが出力される場合（データ読出し）の両方で、それぞれ保証コードを演算し保持する。

【0014】

上位装置にデータを送信する場合、制御部は、内部メモリにデータを格納するときに算出される入力保証コードと、内部メモリからデータを読み出すときに算出される出力保証コードとを比較することにより、正常なデータ送信が行われたか否かを判定することができる。

【0015】

上位装置に送信したデータの一部に障害（パケット未着、データ破壊等）が生じた場合は、送信済データのうち障害の生じた一部分のデータ（部分データ）が再び上位装置に送信される。この場合、制御部は、所望された部分データを通信部を介して上位装置に先に送信した後、内部メモリに格納されているデータ全体の出力保証コードを再計算し、この改めて算出された出力保証コードと入力保証コードとを比較する。両保証コードが適合する場合、制御部は、正常な部分データが送信されたものと判定する。ここで、両保証コードが適合するとは、両保証コードの比較によってデータが正常であることを示す状態を意味し、例えば、両保証コードが一致する場合を含む。

【0016】

このように、要求された部分データを先に送信した後で、この部分データを含むデータ全体について出力保証コードを改めて算出し、この出力保証コードと算出済の入力保証コードとを比較することにより、制御部は、間接的に部分データのデータ保証を行うことができる。これにより、1つのフレームに格納されるデータの全体について入力保証コード及び出力保証コードを算出する保証コード演算部を用いた場合でも、部分データのデータ保証を行うことができる。従って、例えば、チャンネルアダプタと上位装置との間のデータ授受をTCP/IPプロトコルに従って行い、チャンネルアダプタと外部メモリとの間のデータ授受をファイバチャンネルプロトコルに従って行うような場合に有効である。

なお、チャンネルアダプタと外部メモリとの間のプロトコルは、ファイバチャンネルプロトコルに限定されない。また、チャンネルアダプタと上位装置との間のプロトコルは、TCP/IPに限定されない。

【0017】

制御部は、部分データを上位装置に再送信する場合に、（１）通信部により部分データを上位装置に送信させ、（２）データを内部メモリから読み出して通信部にダミー転送し、（３）ダミー転送されるデータについての出力保証コードを保証コード演算部により再計算させ、（４）再計算された出力保証コードとデータを内部メモリに記憶させる際に算出された入力保証コードとを比較し、（５）再計算された出力保証コードと入力保証コードとが一致する場合は、部分データの送信が正常に行われた旨を上位装置に通知し、再計算された出力保証コードと入力保証コードとが不一致の場合は、部分データの送信が正常に行われなかった旨を上位装置に通知する、ことができる。

【0018】

部分データを再送信する場合、この部分データを含むデータの全体を内部メモリから読み出して通信部にダミー転送する。ここで、ダミー転送とは、出力保証コードを算出するために、内部メモリのデータを形式的に通信部に転送することを意味し、通信部に入力されたデータは上位装置に送信されることなく破棄される。

【0019】

あるいは、制御部は、部分データを上位装置に再送信する場合に、（１）部分データを内部メモリから読み出して通信部に転送することにより、部分データを上位装置に送信さ

せ、(2) 保証コード演算部により、内部メモリから読み出された部分データについて部分出力保証コードを算出させ、(3) 部分出力保証コードとデータを内部メモリに記憶させる際に算出された入力保証コードとが不一致であることを確認し、(4) データを内部メモリから読み出して通信部にダミー転送し、(5) ダミー転送されるデータについての出力保証コードを保証コード演算部により再計算させ、(6) 再計算された出力保証コードとデータを内部メモリに記憶させる際に算出された入力保証コードとを比較し、(7) 再計算された出力保証コードと入力保証コードとが一致する場合は、部分データの送信が正常に行われた旨を上位装置に通知し、再計算された出力保証コードと入力保証コードとが不一致の場合は、部分データの送信が正常に行われなかった旨を上位装置に通知する、ことができる。

【0020】

正常なデータ処理が行われている場合、部分データに基づく部分出力保証コードと、データ全体に基づく入力保証コードとは、一致しない。従って、部分出力保証コードと入力保証コードとを比較することにより、保証コード演算部が正常に稼働していることを確認する。

【発明を実施するための最良の形態】

【0021】

以下、図1～図7に基づき、本発明の実施の形態を説明する。

【0022】

本実施の形態には、ディスクアレイ装置に用いられ、上位装置との間のデータ授受を制御するチャネルアダプタの制御方法が開示されている。この制御方法は、上位装置への送信済データの部分データを上位装置に再送信する場合は、内部メモリから部分データを読み出して上位装置に送信させ、データの全体について算出された出力保証コードと内部メモリにデータを入力する際に算出された入力保証コードとを比較し、これら両保証コードが適合する場合に部分データの送信が正常に行われたと判定させる。

【0023】

また、本実施の形態には、チャネルアダプタの別の制御方法も開示される。この制御方法は、外部メモリからのデータを内部メモリに記憶させる第1ステップと、内部メモリにデータを入力する際に、データの全体について入力保証コードを算出し保持する第2ステップと、内部メモリからデータを読み出す第3ステップと、読み出されたデータの全体について出力保証コードを算出し保持する第4ステップと、読み出されたデータを上位装置に送信させる第5ステップと、入力保証コードと出力保証コードとを比較することにより、正常に送信が行われたか否かを判定し、上位装置に通知する第6ステップと、上位装置に送信されたデータの部分データについて、上位装置から再送信要求があったか否かを判定する第7ステップと、部分データの再送信要求があった場合は、内部メモリから部分データを読み出して、上位装置に送信させる第8ステップと、内部メモリに記憶されたデータを再度読み出す第9ステップと、再度読み出されたデータの全体について改めて出力保証コードを算出する第10ステップと、改めて算出された出力保証コードと第2ステップで算出された入力保証コードとを比較することにより、正常に送信が行われたか否かを判定し、上位装置に通知する第11ステップと、を含む。

【0024】

本実施の形態では、ディスクアレイ装置が開示されている。このディスクアレイ装置は、上位装置に通信ネットワークを介して接続され、上位装置との間のデータ授受を制御するチャネルアダプタと、データを記憶する記憶デバイスと、記憶デバイスとの間のデータ授受を制御するディスクアダプタと、記憶デバイスから読出したデータまたは上位装置から受信したデータを記憶するキャッシュメモリとを備え、チャネルアダプタは、キャッシュメモリからのデータを記憶する内部メモリと、内部メモリにデータを記憶させる場合に、データについて入力保証コードを算出して保持する入力保証コード演算部と、内部メモリからデータを読み出す場合に、データについて出力保証コードを算出して保持する出力保証コード演算部と、内部メモリから読み出されたデータを上位装置に送信する通信部と

、制御部とを備え、制御部は、データの部分データを上位装置に再送信する場合に、(1) 通信部によって部分データを上位装置に送信させ、(2) 出力保証コード演算部によりデータについて出力保証コードを再計算させ、(3) 再計算された出力保証コードと入力保証コードとを比較し、これら両保証コードが適合する場合に部分データの送信が正常に行われたと判定する。

【実施例 1】

【0025】

図1～図6に基づいて、本発明の第1の実施の形態を説明する。図1は、ディスクアレイ装置10の全体概要を示すブロック図である。ディスクアレイ装置10は、それぞれ後述するように、各チャネルアダプタ(以下、CHAと略記)20と、各ディスクアダプタ(以下、DKAと略記)30と、共有メモリ40と、キャッシュメモリ50と、スイッチ部60と、各ディスクドライブ70とを備えて構成されている。CHA20及びDKA30は、例えば、プロセッサやメモリ等が実装されたプリント基板と、制御プログラムとの協働により実現される。

【0026】

ディスクアレイ装置10は、通信ネットワークCN1を介して、複数のホストコンピュータ1と双方向通信可能に接続されている。ここで、通信ネットワークCN1は、例えば、LAN(Local Area Network)やインターネット等のようなTCP/IPプロトコルに従って双方向のデータ通信を行うネットワークである。なお、複数のCHA20の全てがTCP/IPに基づいてデータ転送を行う必要はない。一部のCHA20は、例えば、SAN、FICON(Fibre Connection:登録商標)、ESCON(Enterprise System Connection:登録商標)、ACONARC(Advanced Connection Architecture:登録商標)、FIBARC(Fibre Connection Architecture:登録商標)等の通信プロトコルに従ってデータ転送を行ってもよい。

【0027】

各ホストコンピュータ1は、例えば、サーバ、パーソナルコンピュータ、ワークステーション、メインフレーム、携帯情報端末等として実現されるものである。例えば、各ホストコンピュータ1は、図外に位置する複数のクライアント端末と別の通信ネットワークを介して接続されている。各ホストコンピュータ1は、例えば、各クライアント端末からの要求に応じて、ディスクアレイ装置10にデータの読み書きを行うことにより、各クライアント端末へのサービスを提供する。

【0028】

CHA20は、ホストコンピュータ1との間のデータ転送を制御するものである。ディスクアレイ装置10には、例えば、4個や8個等のように、複数のCHA20が設けられている。CHA20は、例えば、オープン系用CHA、メインフレーム系用CHA等のように、ホストコンピュータ1の種類に応じて、用意される。本実施例では、各CHA20のうち少なくとも1台は、TCP/IPに基づいてデータ転送を行う。各CHA20は、図2と共に後述するように、ポート部210と、データ転送制御部220と、プロセッサ部230と、ローカルメモリ240とを備えている。

【0029】

各CHA20は、それぞれに接続されたホストコンピュータ1から、データの読み書きを要求するコマンドやデータを受信し、ホストコンピュータ1から受信したコマンドに従って動作する。例えば、CHA20は、ホストコンピュータ1からデータの読出し要求を受信すると、読出しコマンドを共有メモリ40に記憶させる。DKA30は、共有メモリ40を随時参照しており、未処理の読出しコマンドを発見すると、ディスクドライブ70からデータを読み出して、キャッシュメモリ50に記憶させる。CHA20は、キャッシュメモリ50に移されたデータを読み出し、ローカルメモリ240等を経由して、コマンド発行元のホストコンピュータ1に送信する。また例えば、CHA20は、ホストコンピュータ1からデータの書込み要求を受信すると、書込みコマンドを共有メモリ40に記憶させると共に、受信データをキャッシュメモリ50に記憶させる。DKA30は、共有メモリ40に記憶されたコマンドに従って、キャッシュメモリ50に記憶されたデータを所

定のディスクドライブ70に記憶させる。

【0030】

DKA30は、各ディスクドライブ70との間のデータ通信を制御するものである。ディスクアレイ装置10内には、例えば4個や8個等のように複数のDKA30を設けることができる。各DKA30は、各ディスクドライブ70との間のデータ通信を制御するもので、それぞれプロセッサ部と、データ通信部と、ローカルメモリ等を備えている（いずれも不図示）。各DKA30と各ディスクドライブ70とは、例えば、SAN等の通信ネットワークCN2を介して接続されており、ファイバチャネルプロトコルに従ってブロック単位のデータ転送を行う。各DKA30は、ディスクドライブ70の状態を随時監視しており、この監視結果は内部ネットワークCN3を介してSVP2に送信される。

【0031】

ディスクドライブ70は、例えば、ハードディスクドライブ（HDD）や半導体メモリ装置等として実現される。ここで、例えば、4個のディスクドライブ70によって1つのRAIDグループ80を構成することができる。RAIDグループ80とは、例えば、RAID5（RAID5に限定されない）に従って、データの冗長記憶を実現するディスクグループである。各RAIDグループ80により提供される物理的な記憶領域の上には、論理的な記憶領域である論理ボリューム81（LU）を少なくとも1つ以上設定することができる。このLU81には、ホストコンピュータ1からアクセスされるユーザ領域や、制御情報の格納に使用されるシステム領域等が設定される。

【0032】

共有メモリ40は、例えば、不揮発メモリによって構成されており、制御情報や管理情報等を記憶する。キャッシュメモリ50は、「外部メモリ」の一例であって、主としてデータを記憶する。

【0033】

SVP（Service Processor）2は、ディスクアレイ装置10の管理及び監視を行うためのコンピュータ装置である。SVP2は、ディスクアレイ装置10内に設けられた通信ネットワークCN3を介して、各CHA20及び各DKA30等から各種の環境情報や性能情報等を収集する。SVP2が収集する情報としては、例えば、装置構成、電源アラーム、温度アラーム、入出力速度（IOPS）等が挙げられる。通信ネットワークCN3は、例えば、LANとして構成される。システム管理者は、SVP2の提供するユーザインターフェースを介して、RAID構成の設定、各種パッケージ（CHA、DKA、ディスクドライブ等）の閉塞処理等を行うことができる。

【0034】

図2は、CHA20の概略構成を示すブロック図である。CHA20は、ポート部210と、データ転送制御部220と、プロセッサ230と、ローカルメモリ240とを備えている。

【0035】

「通信部」の一例であるポート部210は、トランシーバ211と、プロトコル制御部212とを備えている。トランシーバ211は、プロトコル制御部212からの制御命令に従い、通信ネットワークCN1を介して、ホストコンピュータ1との間のデータ通信を行う。プロトコル制御部212は、プロセッサ230からの制御命令に応じてトランシーバ211の動作を制御することにより、ホストコンピュータ1との間のデータ転送動作を制御する。

【0036】

データ転送制御部220は、キャッシュメモリ50とローカルメモリ240との間のデータ転送と、ローカルメモリ240とポート部210との間のデータ転送とを、それぞれ制御する。データ転送制御部220は、キャッシュメモリ50ーローカルメモリ240間のデータ転送と、ポート部210ーローカルメモリ240間のデータ転送とを、プロセッサ230を介さずに、DMA（Direct Memory Access）転送する。

【0037】

データ転送制御部220は、入力保証コード演算部221と、出力保証コード演算部222とを備えている。入力保証コード演算部221及び出力保証コード演算部222は、「保証コード演算部」の一例である。入力保証コード演算部221は、キャッシュメモリ50からローカルメモリ240に入力されるデータについて、保証コード（例えば、CRC等）を生成し、保持するものである。出力保証コード演算部222は、ローカルメモリ240からポート部210に出力されるデータについて、保証コードを生成し、保持するものである。各保証コード演算部221、222は、それぞれハードウェア回路として構成されているが、これに限らず、保証コード演算機能の全部または一部をコンピュータプログラムによって実現してもよい。

【0038】

入力保証コード演算部221の算出する入力保証コードGD1は、ローカルメモリ240にデータを入力されるときに生成されるものである。出力保証コード演算部222の算出する出力保証コードGD2、GD3は、ローカルメモリ240からデータを出力するときに生成されるものである。従って、入力保証コードを入力側保証コードやデータ読み込み時保証コード等と、出力保証コードを出力側保証コードやデータ読出し時保証コード等と表現することもできる。

【0039】

プロセッサ230は、CHA20の全体動作を制御するもので、ポート部210及びデータ転送制御部220に適宜制御命令を出力する。プロセッサ230は、ポート部210及びデータ転送制御部220と協働して、後述する処理を実行する。

【0040】

ローカルメモリ240は、CHA20内に設けられるもので、データバッファとしての機能を果たすものである。ホストコンピュータ1へデータを送信する場合、キャッシュメモリ50から所定のデータが読み出されて、ローカルメモリ240に格納される。

【0041】

図2に示すように、ホストコンピュータ1がデータRDを要求すると、DKA30は、要求されたデータRDをディスクドライブ70から読み出して、キャッシュメモリ50に格納させる。データ転送制御部220は、キャッシュメモリ50に格納されたデータRDを読み出して、ローカルメモリ240に転送する。また、このとき入力保証コード演算部221により、データRDについて入力保証コードが算出される。

【0042】

ローカルメモリ240に格納されたデータRDは、データ転送制御部220により、ローカルメモリ240から読み出されて、ポート部210に転送される。このとき出力保証コード演算部222により、データRDについて出力保証コードが算出される。

【0043】

ポート部210に転送されたRDは、TCP/IPプロトコルに従ったパケットに分割され、トランシーバ211からホストコンピュータ1に送信される。入力保証コードと出力保証コードとを比較し、両保証コードが一致する場合は、ホストコンピュータ1に送信されたデータが、要求されたデータRDであることを保証することができる。

【0044】

ホストコンピュータ1に送信したデータRDの一部に何らかの障害が発生した場合、ホストコンピュータ1は、障害で正常に受信できなかった部分データRDpの再送信を要求する。この場合、ローカルメモリ240に格納されているデータRDから、要求された部分データRDpが読み出され、ポート部210を介して再送信される。

【0045】

その後、ローカルメモリ240に格納されているデータRDの全体がポート部210にダミー転送され、このダミー転送によって、データRDの出力保証コードが改めて算出される。ポート部210にダミー転送されたデータRDは、外部に送信されることなく破棄される。改めて算出された出力保証コードと、ローカルメモリ240にデータを格納したときに算出された入力保証コードとを比較することにより、先に送信された部分データR

D p が正常なデータであるか否かを（間接的に）保証することができる。

【0046】

図3は、CHA20からホストコンピュータ1に送信されるデータ構造等を示す説明図である。まず、図3(a)に示すように、iSCSIのフレームは、MAC(Media Access Control Address)アドレス等を含むヘッダ300と、IPパケット310と、TCPパケット320と、iSCSI PDU(Protocol Data Unit)330とを備え、PDU330内にSCSIコマンド、データ、SCSIレスポンスが格納されている。即ち、SCSIコマンドやデータ等をTCPパケットにカプセル化することにより、SCSIデータ等をIPネットワークを経由して送受信できるようになっている。

【0047】

図3(b)は、iSCSIプロトコルの階層構造を示す説明図である。階層の下から順番に、データリンク層/物理層、IP層、TCP層、iSCSI層、SCSI層、アプリケーション層が重ねられている。アプリケーション層(ホストコンピュータ上のファイルシステムやCHA)では、ブロック単位でのデータ転送を行う。SCSI層では、SCSIコマンドを用いてデータの送受信を行う。SCSI層では、ターゲットID、L U番号、L B A(Logical Block Address)によりアドレスを特定する。iSCSI層では、iSCSIネームによりアドレスを特定する。TCP層ではポート番号を使用し、IP層ではIPアドレスを用いる。iSCSI層は、SCSI層とTCP層の間に位置し、SCSI層から受け取ったデータ等をカプセル化してiSCSI PDU330を生成し、TCP層に渡す。また、iSCSI層は、TCP層から受け取ったiSCSI PDU330からSCSIコマンドやデータを抽出し、SCSI層に引き渡す。

【0048】

図4は、ディスクアレイ装置の全体動作の概略を示すフローチャートである。まず、ホストコンピュータ1はCHA20にログインし、ログイン認証等を行う(S1)。次に、ホストコンピュータ1は、リードコマンドを送信することにより、CHA20に対してデータの送信を要求する(S2)。CHA20は、ホストコンピュータ1からの要求を受け付けると、共有メモリ40にリードコマンドを記憶させることにより、DKA30に対してデータの読み出しを要求する(S3)。

【0049】

DKA30は、ディスクドライブ70からデータを読み出し、キャッシュメモリ50に記憶させる(S5)。CHA20は、キャッシュメモリ50に記憶されたデータを読み出し(S6)、ホストコンピュータ1に送信する(S7)。ホストコンピュータ1へのデータ送信が完了すると、CHA20は、データ送信が正常に行われたことを示す正常ステータスをホストコンピュータ1に通知する(S8)。

【0050】

なお、要求されたデータが既にキャッシュメモリ50に記憶されている場合は、キャッシュメモリ50内のデータがCHA20に読み込まれる。より詳しくは、ポート部210がデータ読み出しを要求するiSCSIコマンドフレームを受信すると、このコマンドはプロセッサ230に入力され、プロセッサ230によってコマンド内容が解釈される。プロセッサ230は、データ送信に必要なキャッシュ領域をキャッシュメモリ50に確保し、データ転送制御部220にデータの読み込みを指示する。

【0051】

CHA20からのデータ送信が正常に行われた場合であっても、ホストコンピュータ1では一部のデータを読み出せない場合がある。例えば、ネットワークが混雑していたり、ホストコンピュータ1の処理負荷が大きかったりした場合は、一部のパケットが途中で消失したり、あるいはデータ化けを起こすことがある。そこで、ホストコンピュータ1は、正常に読み込むことができなかった一部のデータについて、CHA20に再送信を要求する(S9)。

【0052】

CHA20は、部分データの再送信要求を受信すると、キャッシュメモリ50に残されているデータを読み出し(S10)、要求された部分データのみをホストコンピュータ1

に再送信する（S11）。なお、ローカルメモリ240内にデータが残されている場合は、ローカルメモリ240から部分データを読み出して、ホストコンピュータ1に再送信することもできる。CHA20は、部分データの再送信を完了すると、正常にデータ送信が行われたか否かを判定し、正常ステータスまたは異常ステータスをホストコンピュータ1に送信する（S12）。

【0053】

図5は、CHA20によるデータ転送処理の具体例を示すフローチャートである。図5のフローチャートは、図4中のS7、S8に示す処理に相当する。また、図5に示すフローチャートは、CHA20によって実行されるもので、より詳しくは、プロセッサ230、ポート部210、データ転送制御部220の協働作業により実行される。

【0054】

データ転送制御部220は、プロセッサ230からの指示に基づいて、キャッシュメモリ50に記憶されているデータRDをローカルメモリ240に転送する（S21）。このとき、入力保証コード演算部221は、ローカルメモリ240に格納されるデータRDについて入力保証コードGD1を算出し保持する（S22）。

【0055】

データ転送制御部220は、プロセッサ230からの指示に基づいて、ローカルメモリ240に格納されているデータRDを読み出し、このデータRDをポート部210に転送する（S23）。このとき、出力保証コード演算部222は、ローカルメモリ240から読み出されたデータについて出力保証コードGD3を算出し保持する（S24）。

【0056】

ポート部210に転送されたデータRDは、プロトコル制御部212によりカプセル化処理され、トランシーバ211から通信ネットワークCN1を介してホストコンピュータ1に送信される（S25）。

【0057】

次に、CHA20は、データRDの送信に際して障害が発生したか否かを判定する（S26）。障害とは、例えば、パケットの消失や未着、データエラー等である。障害が発生していない場合、即ちデータRDを全て正常に送信できた場合は（S26:NO）、CHA20は、入力保証コードGD1と出力保証コードGD3とが一致するか否かを比較する（S27）。両保証コードGD1、GD3が一致する場合（S27:YES）、CHA20は、ホストコンピュータ1に正常ステータスを送信する（S28）。両保証コードGD1、GD3が不一致の場合（S27:NO）、CHA20は、ホストコンピュータ1に異常ステータスを送信する（S29）。

【0058】

一方、データRDを送信したときに障害発生が検知された場合は（S26:YES）、再送信処理が行われる（S30）。この再送信処理の具体例を図6のフローチャートと共に説明する。再送信処理は、CHA20により実行される。

【0059】

CHA20は、ホストコンピュータ1からの再送信要求を受信すると（S301）、再送信を要求されたデータが、データRDの一部分であるかを判定する（S302）。データRDの一部分の再送信要求ではない場合は（S302:NO）、データRDの全体が要求されている場合である。そこで、図5中のS23に戻り、データRDを再送信する。

【0060】

一方、ホストコンピュータ1からデータRDの一部について再送信を要求された場合（S302:YES）、CHA20は、要求された部分データRDpをローカルメモリ240からポート部210のプロトコル制御部212に転送させる（S303）。部分データRDpは、カプセル化処理されてトランシーバ211からホストコンピュータ1に送信される（S304）。このとき、出力保証コード演算部222は、部分データRDpについて出力保証コードGD2を算出し保持する（S305）。

【0061】

CHA20は、部分データRDpについて得られた出力保証コード（部分出力保証コード）GD2と入力保証コードGD1とが一致するか否かを判定する（S306）。入力保証コードGD1は、データRDをローカルメモリ240に格納したときに生成されたものであるから、CHA内部のデータ転送等が正常な場合は、入力保証コードGD1と部分出力保証コードGD2とは一致しない。入力保証コードGD1と部分出力保証コードGD2とが一致した場合（S306:NO）、エラー処理が行われる（S307）。このエラー処理では、例えば、異常ステータスがホストコンピュータ1に送信される。

【0062】

入力保証コードGD1と部分出力保証コードGD2とが一致しない場合（S306:YES）、CHA内部のデータ転送が正常に行われており、また、データRDの一部をS304で送信したことが推定される。次に、CHA20は、ローカルメモリ240に格納されているデータRDの全体を、ポート部210のプロトコル制御部212にダミー転送させる（S308）。プロトコル制御部212は、ダミー転送されたデータRDを受信するものの、カプセル化処理等は行わずに直ちに破棄する。ローカルメモリ240からポート部210へのダミー転送によって、出力保証コード演算部222は、データRDに関する出力保証コードGD3を改めて算出し保持する（S309）。

【0063】

CHA20は、再度算出された出力保証コードGD3と既に作成済みの入力保証コードGD1とが一致するか否かを判定する（S310）。両保証コードGD1、GD3が一致する場合（S310:YES）、CHA20は、S304で送信した部分データRDpが正常なデータであると判定し、正常な再送信が行われたことを示す正常ステータスをホストコンピュータ1に送信する（S311）。両保証コードGD1、GD3が不一致の場合（S310:NO）、CHA20は、部分データRDpが異常データであることを示す異常ステータスを送信する。

【0064】

本実施例によれば、データRDの一部を構成する部分データRDpについて再送信がホストコンピュータ1から要求された場合でも、部分データRDpが正常なデータであることを間接的に証明してデータ保証を行うことができる。従って、ファイバチャネルプロトコルのように、基本的に、送信済データの一部分についてはデータ保証を行えない環境下であっても、部分データのデータ保証を行うことができる。これにより、ファイバチャネルプロトコル用に開発されたCHA20をIPネットワークに対応させながら、データ保証を行うことができる。また、本実施例では、ポート部210にデータバッファを備えない構成においても、部分データRDpの再送信及びデータ保証を行うことができる。

【実施例2】

【0065】

図7は、第1実施例の変形例に相当するCHA20の概略構成を示す。本実施例の特徴は、入力保証コード及び出力保証コードの両方を1つの保証コード演算部223によって算出し保持するようにした点にある。

【0066】

なお、本発明は、上述した各実施の形態に限定されない。当業者であれば、本発明の範囲内で、種々の追加や変更等を行うことができる。例えば、保証コードとしてはCRCを例示したが、これに限らず、種々の保証コードを採用することができる。

【図面の簡単な説明】

【0067】

【図1】本発明の実施形態に係わるディスクアレイ装置の全体構成を示すブロック図である。

【図2】CHAの概略構成を示すブロック図である。

【図3】（a）はiSCSIのフレーム構造を、（b）はiSCSIの階層構造をそれぞれ示す説明図である。

【図4】ディスクアレイ装置の全体動作の概要を示すフローチャートである。

【図 5】 C H A によるデータ転送処理の概要を示すフローチャートである。

【図 6】 データを再送信する場合の処理概要を示すフローチャートである。

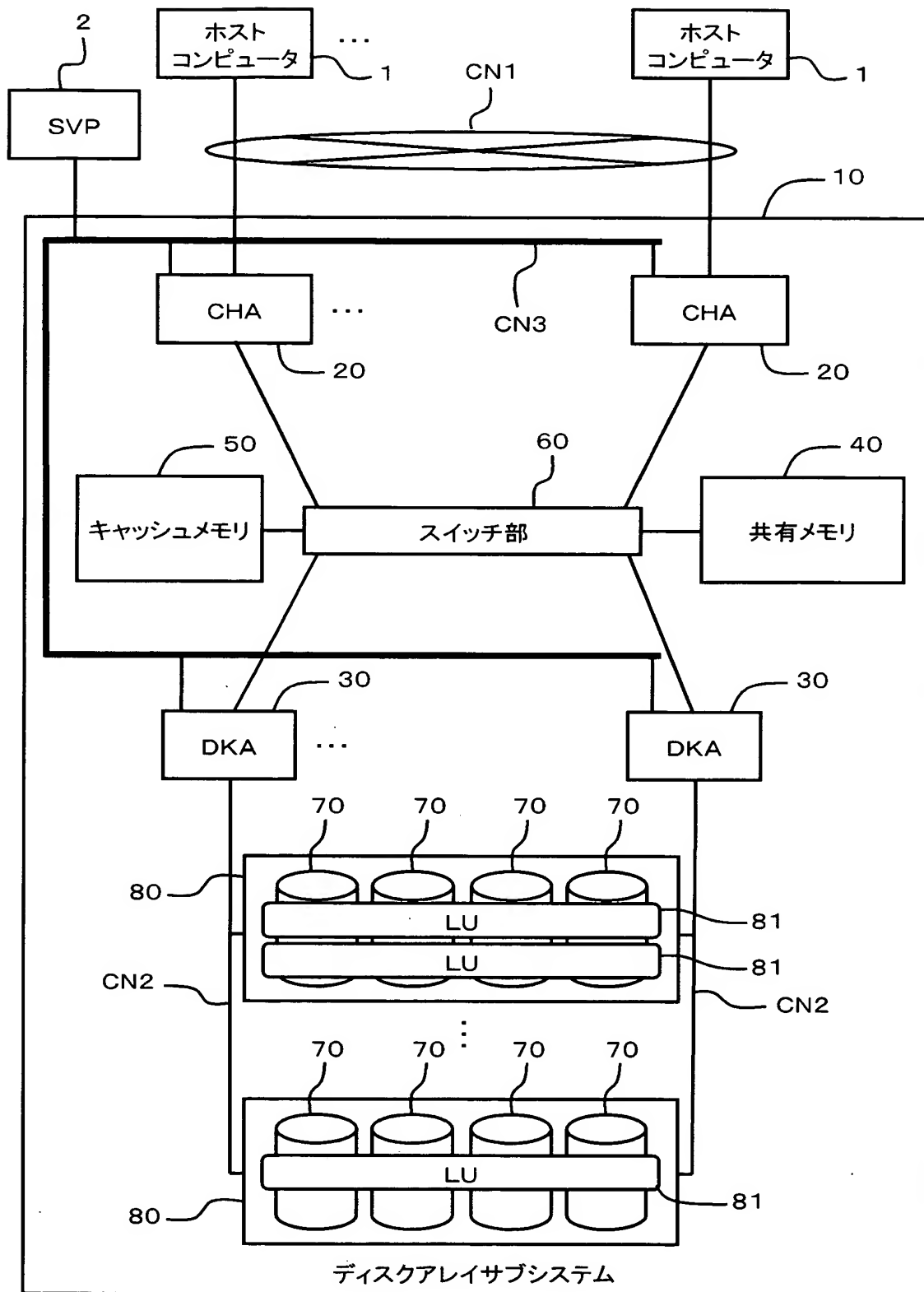
【図 7】 変形例に係わる C H A の概略構成を示すブロック図である。

【符号の説明】

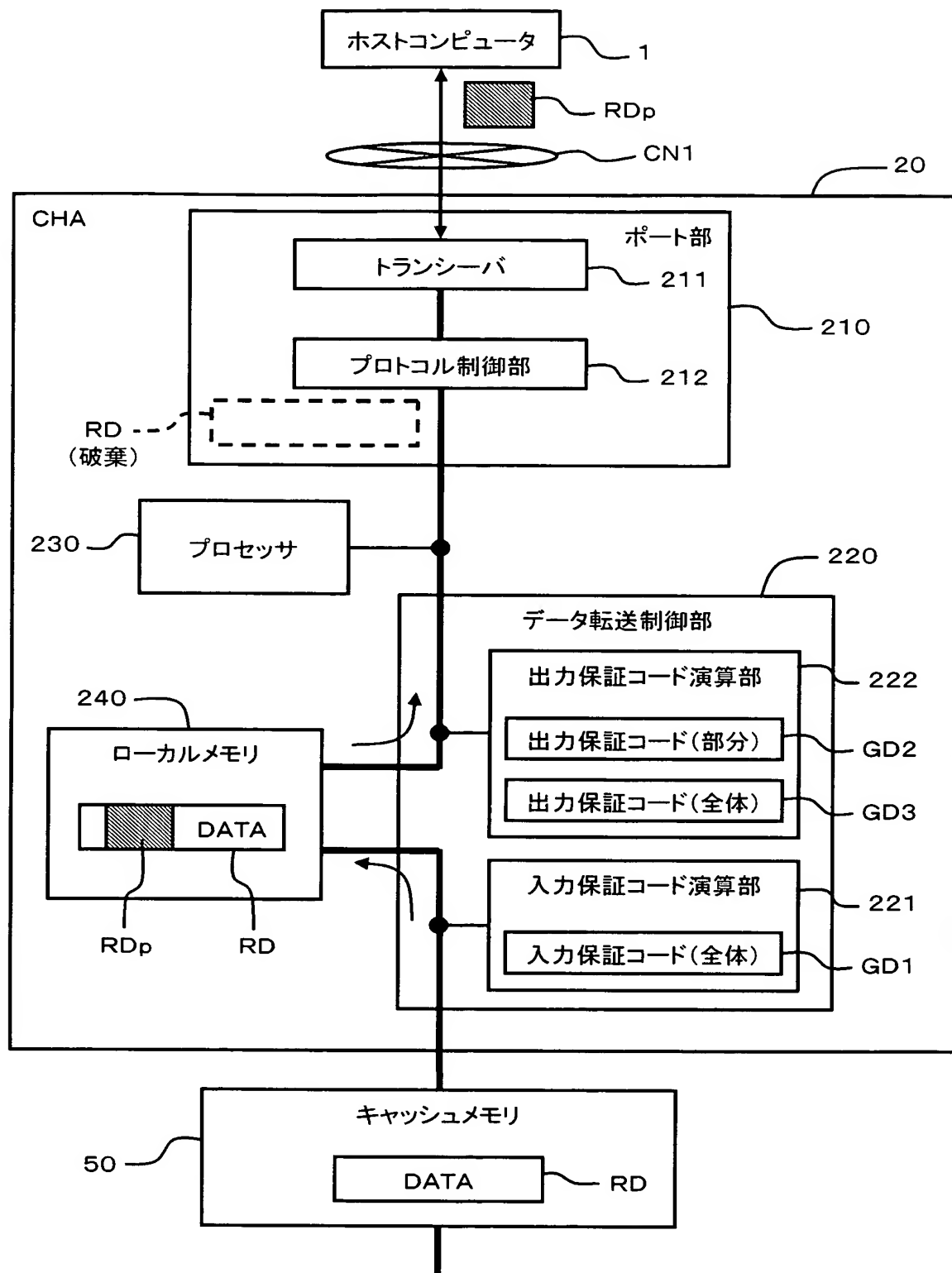
【 0 0 6 8 】

1…ホストコンピュータ、2…SVP、10…ディスクアレイ装置、20…CHA、30…DKA、40…共有メモリ、50…キャッシュメモリ、60…スイッチ部、70…ディスクドライブ、80…RAIDグループ、81…論理ボリューム、210…ポート部、211…トランシーバ、212…プロトコル制御部、220…データ転送制御部、221…入力保証コード演算部、222…出力保証コード演算部、223…保証コード演算部、230…プロセッサ、240…ローカルメモリ、CN1～CN3…通信ネットワーク、GD1…入力保証コード、GD2…部分出力保証コード、GD3…出力保証コード、RD…データ、RDp…部分データ

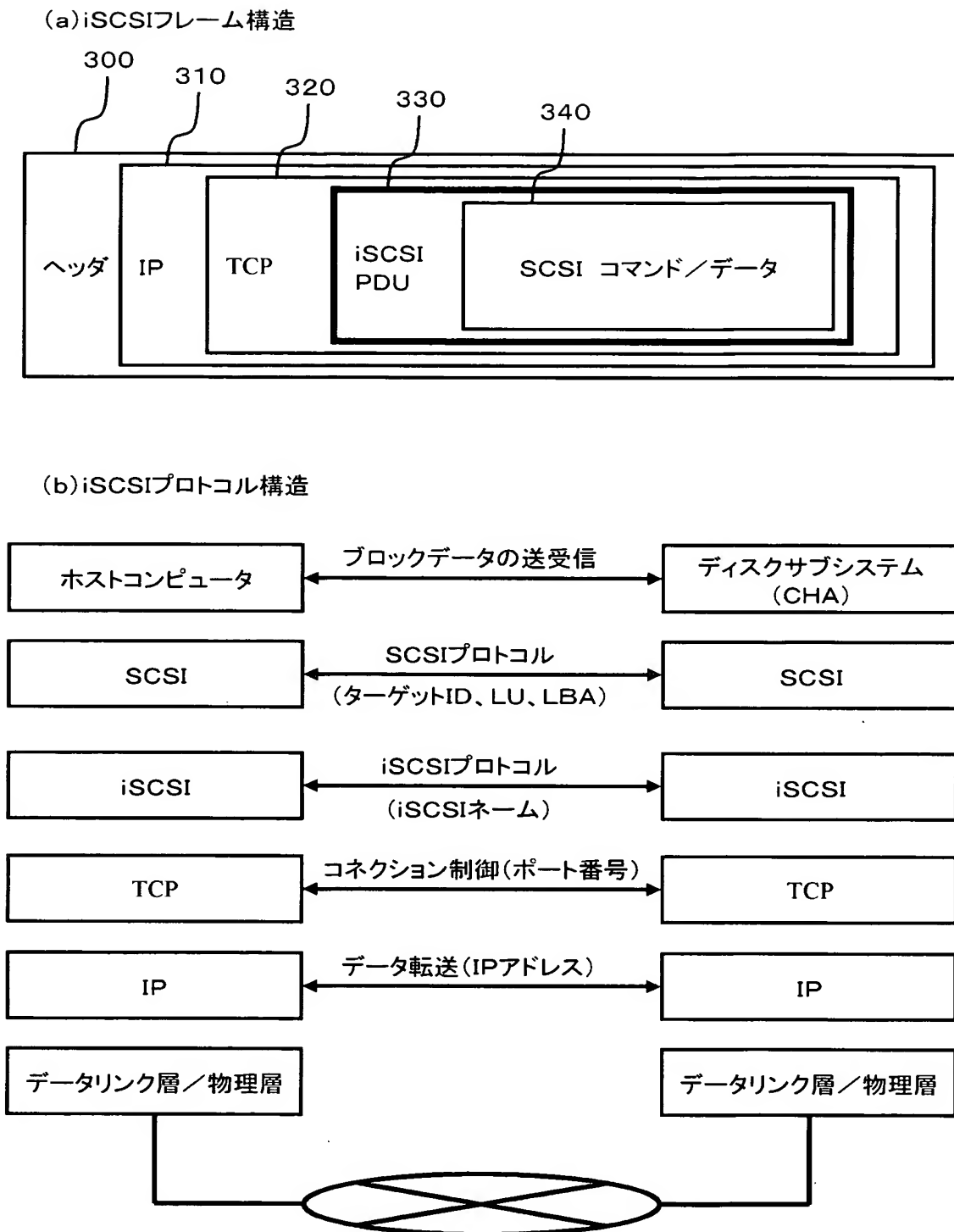
【書類名】 図面
【図 1】



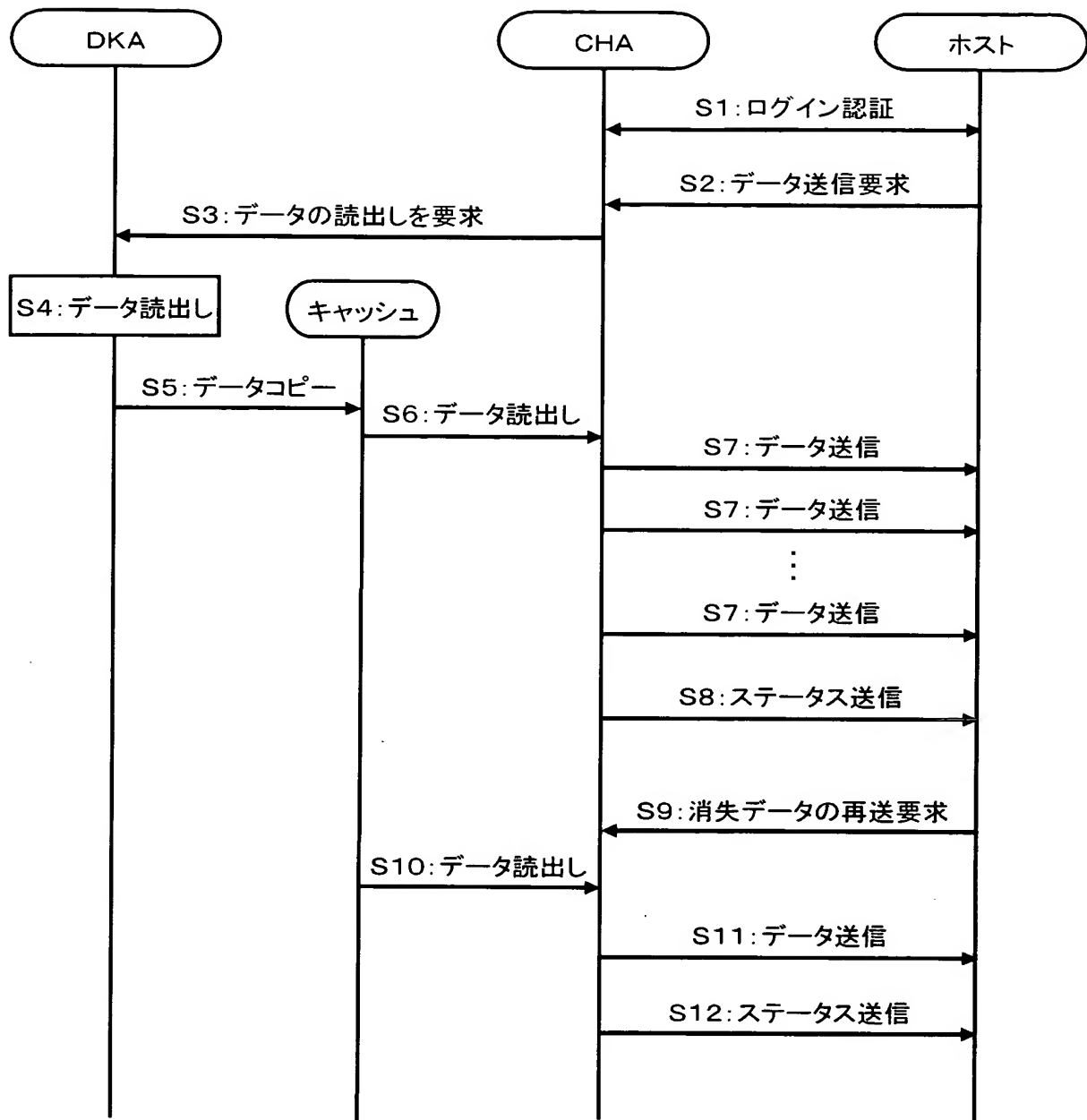
【図 2】



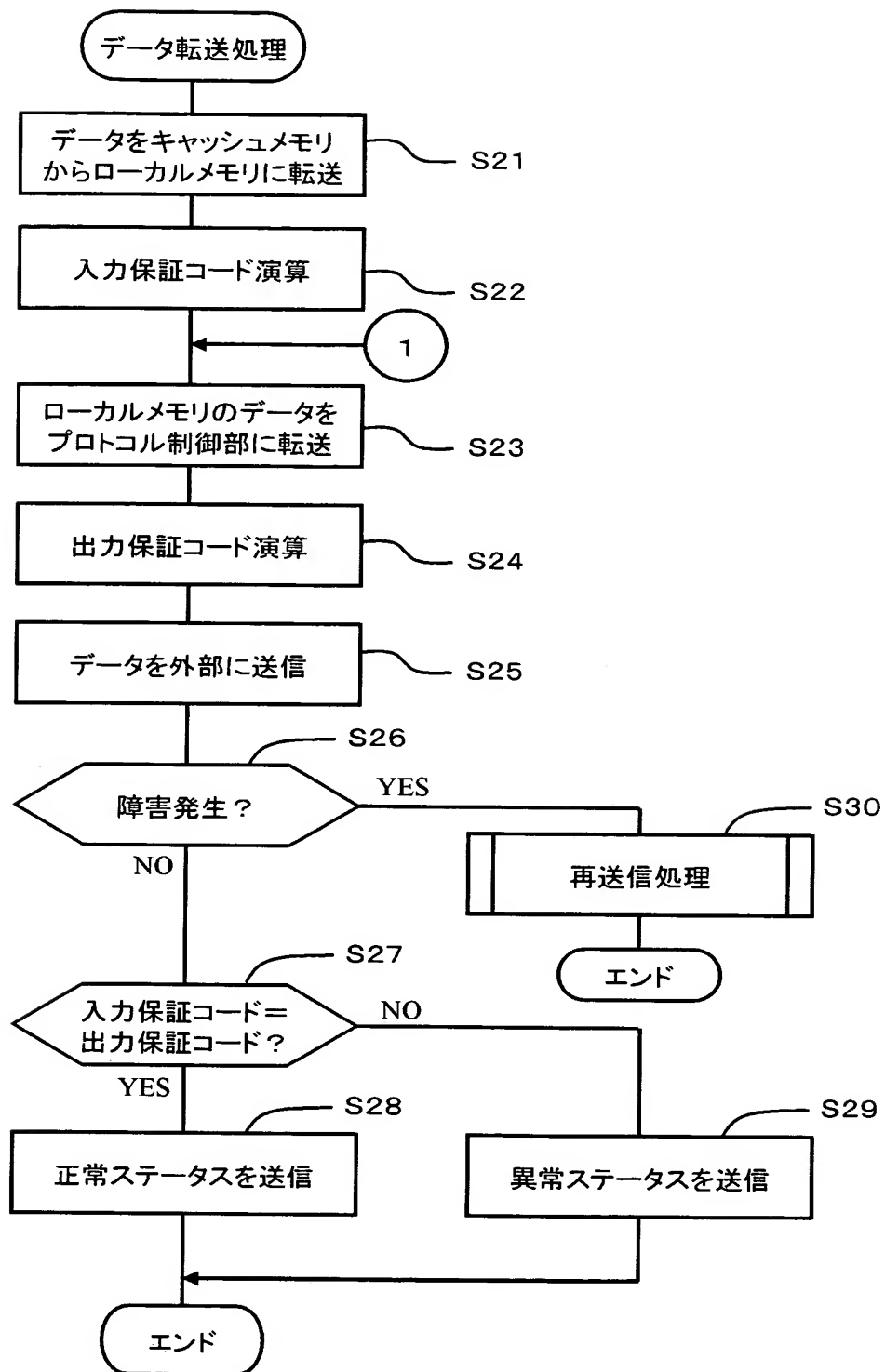
【図 3】



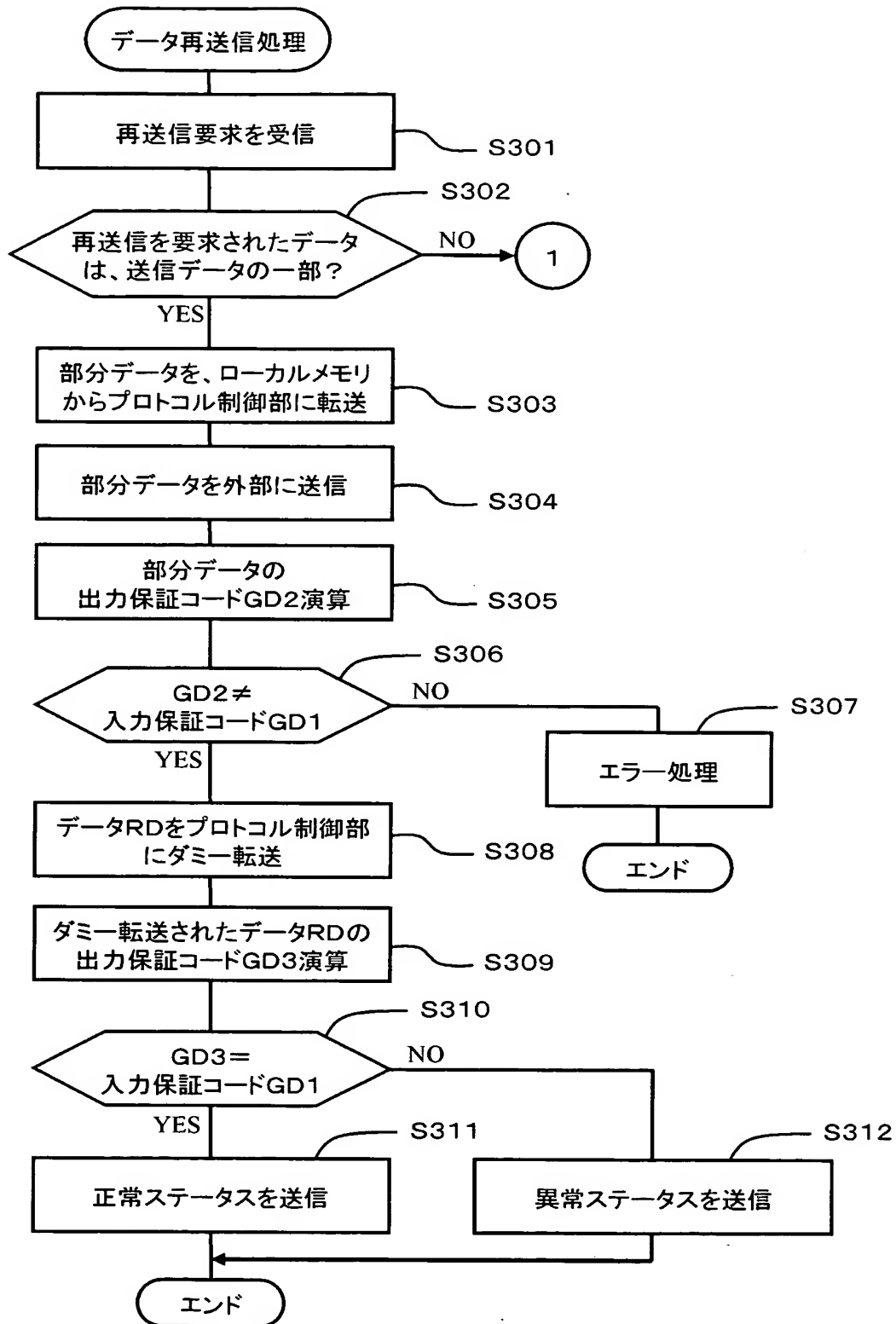
【図 4】



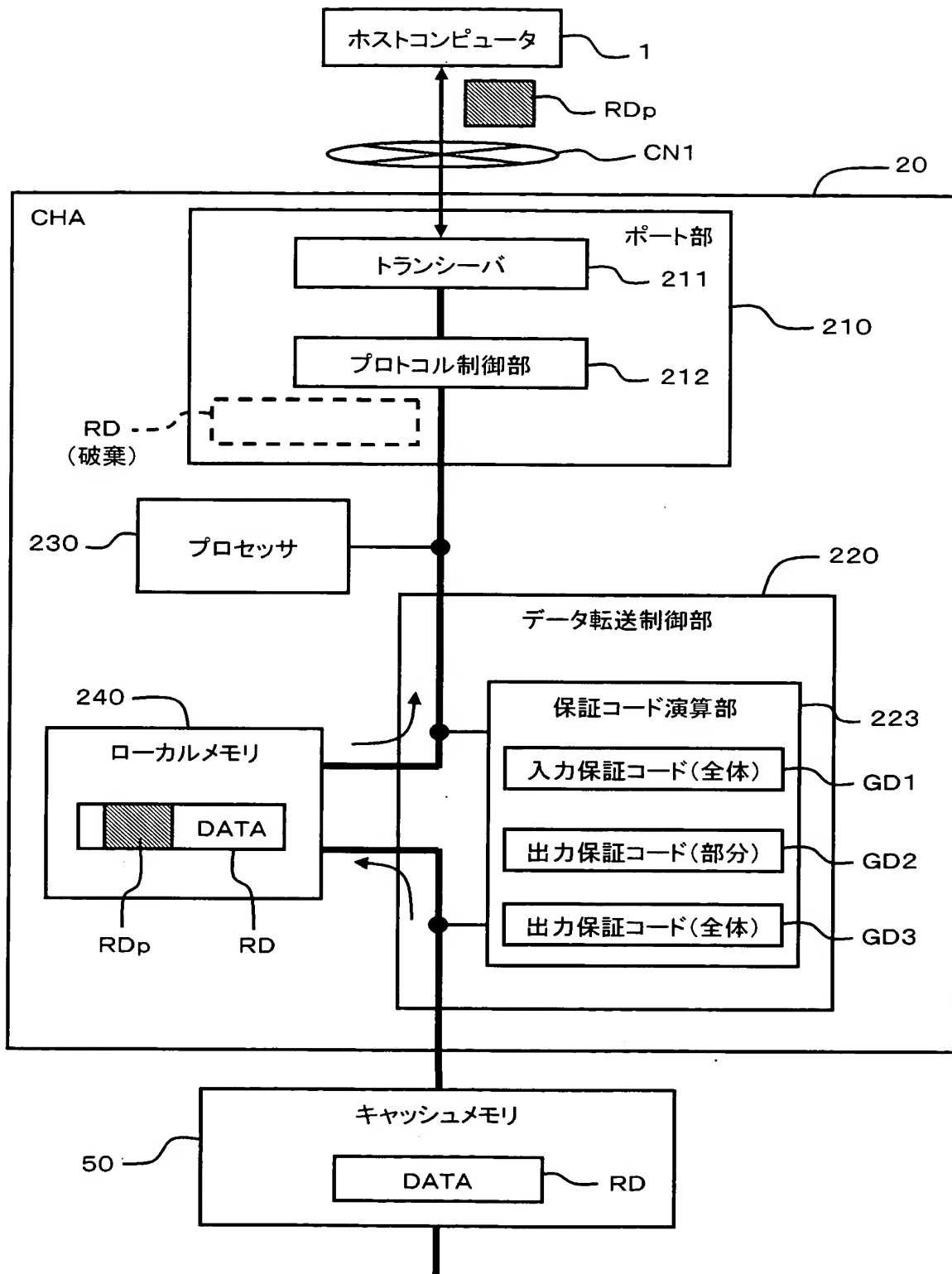
【図 5】



【図 6】



【図 7】



【書類名】 要約書**【要約】**

【課題】 I P ネットワークを介してデータの一部を再送信でき、部分データのデータ保証を行えるようにすること。

【解決手段】 C H A 2 0 は、データ R D をホストコンピュータ 1 に送信する。C H A 2 0 は、データ R D をキャッシュメモリ 5 0 からローカルメモリ 2 4 0 に記憶させる際に入力保証コード G D 1 を算出する。C H A 2 0 は、ローカルメモリ 2 4 0 からデータ R D を読み出してポート部 2 1 0 に転送する際に出力保証コード G D 3 を生成する。データ R D の送信後に、部分データ R D p の再送信を要求された場合、C H A 2 0 は、部分データ R D p をホストコンピュータ 1 に送信する。この後で、C H A 2 0 は、データ R D の出力保証コード G D 3 を再度算出し、先に算出された入力保証コード G D 1 と比較する。両コード G D 1 , G D 3 が一致する場合、正常にデータ送信が行われたことが保証される。

【選択図】 図 2

認定・付加情報

特許出願の番号	特願 2 0 0 3 - 3 9 3 6 4 7
受付番号	5 0 3 0 1 9 3 3 7 1 6
書類名	特許願
担当官	第七担当上席 0 0 9 6
作成日	平成 1 5 年 1 1 月 2 6 日

< 認定情報・付加情報 >

【提出日】 平成15年11月25日

特願 2 0 0 3 - 3 9 3 6 4 7

出 願 人 履 歴 情 報

識別番号 [0 0 0 0 0 5 1 0 8]

1. 変更年月日	1 9 9 0 年 8 月 3 1 日
[変更理由]	新規登録
住 所	東京都千代田区神田駿河台 4 丁目 6 番地
氏 名	株式会社日立製作所